Using Role Determination and Expert Mining in the Enterprise Environment

Jing Yao, Jun Xu, Junyu Niu,
Computer Science and Engineering Dep. of Fudan University,
Shanghai, China 200433

Abstract

In real world, expert search is not just only name matching. Since each kind of people has their own features, we try two methods to judge whether the person we have found is more likely to be an expert. One method is to determine the role of a person by the context of the pages; the other is to judge the authority of a person by the forms of pages where he appears considering the structure of the Intranet. The evaluation results show these new methodologies have been helpful to improve the performance of the expert search on TREC 08 queries.

1. Introduction

There are two traditional categories of solutions we used to adopt for expert search: either looking for the specific topic in every expert's related document, or looking into every document which contains information on the topic and trying finding if there is any expert name appears.[1] In real world, expert search is not just only name matching. Since each kind of people has their own features, we try two methods to judge whether the person we have found is more likely to be an expert. One method is to determine the role of a person by the context of the pages; the other is to judge the authority of a person by the forms of pages where he appears considering the structure of the Intranet. So as for the expert search task this year, we evaluate the candidates from 3 aspects: relevant page retrieval, role determination, and expert page mining. The system architecture is illustrated in figure1.

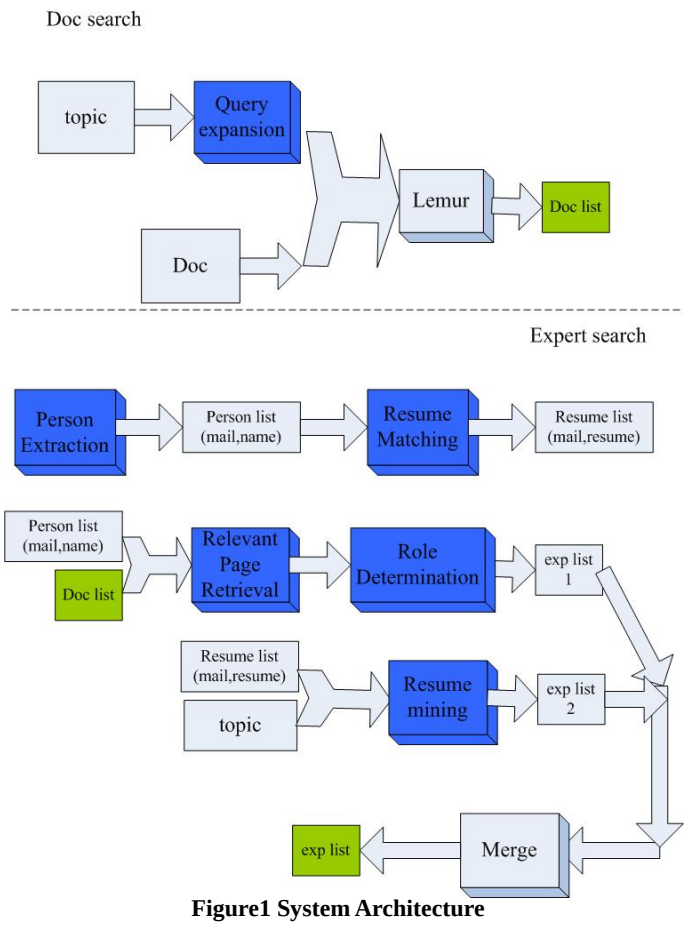

Figure1 System Architecture

Relevant page retrieval is the foundation of our system. Traditional information retrieval model are used at this stage. Each

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2008		**00-00-2008 to 00-00-2008**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Using Role Determination and Expert Mining in the Enterprise Environment	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Fudan University,Computer Science and Engineering Dep,Shanghai, China 200433,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Seventeenth Text REtrieval Conference (TREC 2008) held in Gaithersburg, Maryland, November 18-21, 2008. The conference was co-sponsored bythe National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA).

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**6**	

candidate is assigned a score decided by the relevance between topics and pages, and the relevance between topic-related pages and certain candidate. Thus we could obtain a roughly ranked expert list with supporting page lists for each individual, and then we adjust the expert lists in the next two stages.

At the role determination stage, the role of the candidate could be evaluated by analyzing the context where the candidate appears. This can be achieved by applying statistic approaches to last years' results to discover in what kind of contexts a candidate is more likely to be an expert. For example, if the context indicates that the candidate's role is a team leader, it would be more convincible that this candidate was an expert. Thus we give the candidate a higher score for the certain topic.

As for the expert page mining stage, since Intranet is usually much more structured than Internet, certain kinds of pages in Intranet usually has some definite functions, such as expert introduction, project overview and etc. These pages with specific functions could be retrieved by the forms of the pages or even just by the forms of their URLs. At this stage, we just simply filter the pages by the forms of URLs to retrieve those expert resumes, and candidates who has a resume page relevant to the topic is considered to be more authoritative in that field. This approach could also be extended by deeper mining such kind of resume pages, and since these pages is highly identical in forms, it is more feasible to discover valuable information from them. The following pages

2. Query Expansion

In order to take full advantage of the description part of each query, we assume that the less popular words in the description are more likely to provide valuable information for retrieval than common ones. For this year we used the concept “index of familiarity” in Wordnet to do this job. This idea is first proposed in[2]. The idea is that the familiarity of a word is not determined by its appearance frequency, but is decided by the number of meanings it have. In another word, the more semantic meanings the word has, the more popular the word is. Thus, we use the Wordnet API to automatically find out those words with lower familiarity. Empirically, we assume that proper nouns are more valuable for the query than other parts of speech. Thus, we define two thresholds α, β which denotes the limits of the number of noun meanings and the number of all meanings a phrase may have. If the phrase fulfills following condition:

$$\text{NounMeanings}(word) < \alpha \quad (1)$$

And

$$\text{AllMeanings}(word) < \beta \quad (2)$$

It is then considered valuable to be added to the expanded query.

3. Expert Search Model

3.1 Relevant Page Retrieval

To find experts on a specific query, the problem can be stated as how probably the candidate is the expert in this specified field. Although there is no predefined candidate list given by the organization, we can recognize emails using the pattern “firstname.lastname@csiro.au” as

expert identifiers and found candidates' full names in the context of emails.

We consider the expert as a mixture of documents relevant to the query topic, thus we estimate the probability as follows:

$$p_D(c \mid q) \propto \sum_{d \in D} p(c \mid d)\, p(d \mid q) \quad (3)$$

Where D is the document repository, $p(c \mid d)$ denotes the relationship between the candidate and the document on the topic t and $p(d \mid q)$ denotes the relevance of the document to the given topic. For the $p(d \mid q)$ can be got by search engine like Lucene, Lemur which use classical information retrieval model, we focus our research on evaluation of the relationship between candidates and documents.

3.2 Role Determination

3.2.1 Candidate – Document Relation

Basically we can get $p(c \mid d)$ by the language model which takes the frequency of the candidate's appearance into first consideration.

$$p(c_i \mid d) = \frac{tf(c_i \mid d)}{\sum_{c_j \in C} tf(c_j \mid d)} \quad (4)$$

The presence of the candidate in the document can in several ways besides email and full name. If he appears as an author with his book in the page, the format might be f. lastname, which f. is the abbreviation of the first name of the candidate. The formats we used to detect in document are listed as followed:

firstname lastname firstname.lastname f.lasname
firsname.l lastname, firstname lastname, f

The candidate may also appear in the format like Mr lastname, but barely last name match brought too much noise in our experiment.

But it is not reasonable to evaluate a candidate's expertise just for his high frequency in the document even though the document is highly relevant to the topic. In some situations, the candidate may appear only once, but is more important than other candidates the documents. For example, a web page describes a technical conference where all the attendants' names appear in the page. Obviously the organizers of the conference are more responsible for the page than other candidates. But the basic model cannot reflect the situation. Let's look back the manual way we judge whether a candidate is an expert. We think it depends on the documents in which the candidate appears and the role he acts in the document. We are most likely to judge Michael Robertson is an expert on CSIRO sustainable ecosystem for Michael is a doctor and a scientist. The position of the candidate is the main evidence to prove his expertise. We can list several words of this kind: contact, scientist, researcher, analyst and so on. We call these role determination words.

Figure 2 web page fragment

We take the correct answers of the topics for the last year task as the training data. We extract a 100-word window of the experts' name and make these words as a candidate role determination word list. Then we filter the list by two rules: the part of speech the word and how common the word is. We only remain nouns for they always give more information about who the candidate is. We calculate the word frequency of the words for each topic. If a candidate word only appears in few topics, we think the word is too specific and remove it from the list. At last we get a small word list with the frequency. We divide the words to three levels according to the frequency. A level shows the candidate is highly likely an expert and C level shows the candidate maybe an expert, but it is not certain. We give each expert a role score according to which level of words and the number of words appearing in the context of the candidate's name. A level is 0.6, B level is 0.3, and C level is 0.1.

$$role(c_i \mid d) = \max(1, (0.6 * num(wordA, d) + 0.3 * num(wordB, d) + 0.1 * num(wordC, d)) \quad (5)$$

Now we improve the equation (4) as:

$$p(c_i \mid d) = \max(\frac{tf(c_i \mid d)}{\sum_{c_j \in C} tf(c_j \mid d)}, role(c_i, d)) \quad (6)$$

3.2.2 Candidate - Topic Relation

The purpose of expert search is to find experts on the given query. Although we improve the candidate-document relation by the role mining, the candidate will still get a high score when he appears on quite a lot pages although his roles on these pages are not important at all. As we hope we can turn to the candidates who are really responsible for the research on the specific field, we judged the candidate's role on the query. For example, if the candidate's role in the document is a team leader and the document is a project home page of the query, then we gave the candidate a higher score on the query.

Equation (3) can be rewritten as:

$$ScoreofRoleDetermination(c, q) = \alpha \sum_{d \in D} p(c \mid d) p(d \mid q) \quad (7)$$

Where α evaluates the responsibility of the candidate to the query field.

3.3 Expert Page Mining

Expert search is much more often used in intranet other than Internet. A distinct difference between intranet and internet is that intranet is much more structured and much more specific in certain fields. Thus we could use this characteristic to improve the expert retrieval. The structured

information can be utilized in two ways:

(1) The structure of the intranet infers the potential function and importance of different pages at different positions of the structure. For instance, the higher level a page is located, the more general information it may contain. On the other hand, the pages which are at the deeper levels of the structure may provide more details of certain event. And pages at different positions also provide different functions. As for expert search, those pages under catalog "http://www.csiro.au/people" may be quite valuable since they serve as resume pages of some important staff in the organization.

(2) As the structure of the intranet is well-regulated, the pages which serve the similar function are quite identity in forms. Using CSIRO as an example, the various project overview pages or resume pages have quite similar forms. Thus, it is more convenient for deep mining these kinds of pages since we can identify what kind of information each part of the page contains. Various kinds of approaches can be applied for deeper page mining at this point. Such as VIPS [3], by which we can segment the pages by its forms, and get structured information from them.

Thus, in addition to the former two scores we have, we get the third part of score to rank the experts. We predefine a set of templates of pages which are more likely to introduce an expert. Then the score is determined by two factors. One is the location of the considering the whole intranet structure, which is denoted as page Location(page), the other is the relevance between the page and expert which can be calculated by mining the pages upon its forms information, which is denoted as Mining(page, expert). So the score of expert mining is defined as:

$$ScoreofExpertMining(c,q) = \sum_{d \in D} p(d \mid q) \cdot f(Location(d), Mining(d,c)) \quad (8)$$

Finally, we integrate formula (7) and formula (8) to get the final ranking score of a candidate:

$$ExpertRankingScore(c,q) = ScoreofRoleDeteminaion(c,q) \times (1 + \log(ScoreofExpertMinging(c,q))) \quad (9)$$

4. Results and Analysis

We have submitted four runs for the expert search task. FDUExpBase uses the basic expert search model which did not include the role determination phase and the expert page mining phase. FDUExpRole adopts the role determination method and FDUExpRes adopts the expert page mining method respectively. And FDURoleRes include the both two methods.

Table 1 and table 2 show the experiment results of the four runs. From the tables we can see that the role determination strategy and the expert page mining strategy have helped to improve the performance of the expert search. FDURoleRes performs the best on MAP, while FDUExpRole outperforms the other runs on R-prec, P@5 and P@10.

However, the integration of the two strategies has not made as much improvement as we had expected. The possible reason is that we have adjusted the parameters according to the experiment on TREC 07 queries and they are not quite appropriate for the new queries. Thus, the integration mechanism and the two methods themselves need to be further evaluated and investigated in our future work.

	MAP	R-prec	P@5	p@10
FDUExpBase	0.3720	0.3502	0.4436	0.3436
FDUExpRole	0.4112	**0.3985**	**0.4691**	**0.3582**
FDUExpRes	0.3815	0.3601	0.4291	0.3491
FDURoleRes	**0.4114**	0.3943	0.4618	0.3509

Table 1 Results of Four Runs

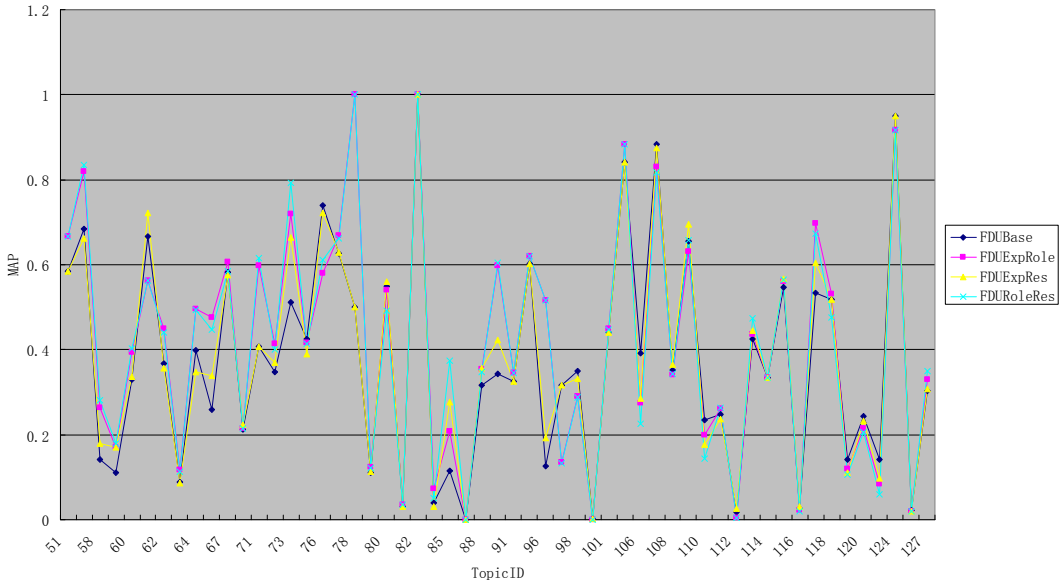

Table 2 Map of Four Runs on the Different Topics

5. Future Work

Compared with Internet, Intranet is much better organized. Analyzing the feature of the certain entities on intranet and preprocessing these useful information in a suitable way could significantly improve the user experience of retrieval. At the moment when we regard “expert” as a kind of entity, we have applied role determination and expert page mining on it and these methods has been proved to be effective. In future we will focus on further investigation on entity finding issues, such as how to describe the features of an entity and what detail level of the description is suitable for entity searching.

References

[1] K. Balog, L. Azzopardi, and M. de Rijke. “Formal models for expert finding in enterprise corpora”, In SIGIR ’06: Proceedings of the 29th annual international ACM SIGIR conference, 2006.

[2] Richard Beckwith, George A. Miller, Randee Tengi, Design and Implementation of the WordNet Lexical Database and Searching Software, available at http://santana.uni-muenster.de/Seminars/WordNetHS_WS02/Texte/Beckwith93DesignWordNet.pdf

[3] Deng Cai, Shipeng Yu, Ji-Rong Wen, Wei-Ying Ma, VIPS: a Vision-based Page Segmentation Algorithm, available at ftp://ftp.research.microsoft.com/pub/tr/tr-2003-79.pdf